

Algorithmic Gender Prediction Is Illegitimate, But Gender Imputation Can Yield Valid Measurements

Evan Dong and Angelina Wang

Cornell University
edong.cs@cornell.edu, angelina.wang@cornell.edu

Abstract

Machine learning ethics researchers, critical HCI scholars, and interdisciplinary academics have argued that algorithmically predicting gender is *wrong*. At the same time, others in the same research communities rely on predicted gender labels to study gender disparities and develop algorithmic fairness techniques. Taking both of these perspectives in good faith pushes us to believe that algorithmic gender prediction simultaneously cannot be trusted and at the same time can identify genuine gender disparities. How do we reconcile these two seemingly contradictory intuitions? We draw a distinction between two questions: whether algorithmic gender prediction is wrong, being *illegitimate* and thereby contributing to harm; and whether algorithmic gender prediction is wrong, being *invalid* and thereby producing unusable measurements. Our analysis translates different arguments against gender prediction into these terms of legitimacy and validity and shows how gender imputation applied for fairness purposes can be both illegitimate and still yield valid disparity measurements. We clarify this bind by drawing upon transfeminist literature to distinguish two forms of sexism: sexism that targets women and femininity, and sexism that targets transgender and nonbinary people. While gender imputation can produce valid measurements for the former, it is illegitimate and harmful for the latter. We argue that practitioners should deploy gender imputation only when it would achieve anti-discrimination benefits that cannot be achieved through other reasonable means, while harms are minimized to the extent possible. We examine how this tension plays out through three case studies: auditing gender bias in generative image models, measuring gender disparities in film, and imputing gender from personal names. By disentangling both questions of legitimacy from questions of validity, as well as these two different forms of sexism, we show how debates over gender prediction have often conflated distinct concerns, obscuring both the settings in which gender imputation can support fairness efforts and the harms towards transgender and nonbinary people that it fundamentally cannot capture. We conclude by recommending the development of more inclusive methods that address all kinds of sexism.

1 Introduction

Assessing algorithmic discrimination generally requires access to demographic data. Consider that quantifying the U.S.

gender pay gap in 2024 — that women earned 85% of what men earned — requires data on workers’ genders (Fry 2025). However, demographic data is not always available. For instance, in the U.S., many financial institutions and government agencies such as the Internal Revenue Service and Patent and Trademark Office are often not legally permitted to collect demographic data such as race and sex (Bogen, Rieke, and Ahmed 2020; King et al. 2023; Kumar, Hines, and Dickerson 2022). Similarly, users may be reluctant to disclose their demographic data for privacy reasons, especially to large institutions they may not trust (Bogen, Rieke, and Ahmed 2020; King et al. 2023; Andrus et al. 2021). In these cases, this demographic data is often *imputed* (i.e., filled in with estimated values via a predictive method); for example, by predicting gender from images or given names, or predicting race from ZIP codes and surnames. Many transgender¹ and critical scholars have opposed this practice when it comes to gender, arguing against prediction because of the direct harm of misgendering individuals and the problematic premises it encodes, which reify gender as binary, static, and predictable from a face or name.

In this work, we focus specifically on the case of predicting gender, rather than other demographics such as race or age. While predicting other demographic attributes faces the same bind of label need and premise critique, and many of our arguments may apply, focusing on gender enables us to engage concretely with the specific critiques posed by, e.g., transgender and nonbinary² advocates, and the gender-specific harms of this premise (Smith and Stewart 1983; Reid 1988; Wang 2025).

The evaluation of gender disparity is thus caught between the desire to detect situations in need of change and the de-

¹“Trans” is a common abbreviation of “transgender,” especially when used as a modifying adjective, e.g., “trans studies,” or a prefix, e.g., “trans-exclusionary.” Like Ovalle et al. (2023), we refer readers to the community-run Nonbinary wiki’s English glossary of gender terms. For convenience, we have transcribed definitions of terms that appear in this paper into App. A.

²Our language in this paper focuses on transgender and nonbinary people. We use these terms with the understanding that, they cover very heterogeneous experiences, and that others under different labels, such as gender non-conforming and genderqueer, are often similarly marginalized, part of similar communities, and hold similar concerns.

sire to dismantle harmful conceptions of static gender. This bind is reflected in conflicting moral intuitions in day-to-day life: consider that many women and nonbinary researchers have experienced the real harms of feeling like the only gender minority in a room seemingly full of men at a computer science conference, even without knowing the actual gender identities of those present. Trans-inclusive principles show us that assuming another person's gender can be wrong; however, these principles do not exist for the sake of invalidating common experiences of gender discrimination.

We navigate this tension by drawing two key distinctions that have previously muddled this discussion. First, we distinguish legitimacy (whether a prediction normatively deserves given social authority) from validity (whether a prediction descriptively measures a concept well). Second, we draw on Julia Serano's analysis of trans-misogyny (Serano 2007) to distinguish traditional sexism (discrimination against women and femininity) and oppositional sexism (discrimination against gender deviance and non-normativity). We argue that these different forms of sexist discrimination are based on different aspects of gender, such as gender identity and perceived gender, that have different legitimacy and validity implications. This allows us to concretely describe which dimensions of sexism that imputation is more suited to measure (i.e., traditional sexism), while also highlighting the dimensions of sexism that are often ignored or actively perpetrated (i.e., oppositional sexism) when overly focusing on the former. This allows us to reconcile this bind and argue that *gender imputation can lead to valid measurement of disparities for traditional sexism, even though its premise is illegitimate and causes harm with respect to oppositional sexism*. We emphasize that this illegitimacy is a serious concern, especially in a time where trans rights are under attack and transgender people are facing intense persecution; our work seeks to navigate this tension without discounting this fact, and in fact to critique works using gender imputation that relegate this harm to a mere acknowledgment or footnote (Devinney, Björklund, and Björklund 2022). As much as we seek to defend imputation for traditional sexism, we also call for action to address the neglect of oppositional sexism.

We lay out our argument in detail before discussing recommendations and case studies. First, in Sec. 2 we provide background on the imputation of demographic data and differentiate imputation as a special case of prediction. Next, in Sec. 3 we examine existing critiques of gender prediction, and analyze which characteristics affect the legitimacy of the prediction. By drawing on trans scholarship, we highlight a range of applications, prediction targets, and gender concepts, demonstrating that all gender prediction is to some degree illegitimate. Then, in Sec. 4 we differentiate these reasons for illegitimacy from the validity of measurements that are relevant to measuring gender discrimination, with some limitations, and in Sec. 5 provide concrete recommendations towards better practices in gender imputation. In Sec. 6 we work through the relevance of different factors in model design and downstream application, sharing three case studies on gender imputation in generative models, film, and personal names, to demonstrate how our argument plays out

in practice. We do not mean to legitimize gender prediction as it is often practiced, which causes severe harm to transgender and nonbinary people — harm that can and should be minimized. We provide recommendations for such harm minimization, and acknowledge only that, in cases where no alternative methods exist, there are benefits of measuring traditional gender discrimination that should not be ignored. In this work, our core contributions are to: (a) work through the bind of algorithmic gender prediction being illegitimate but sometimes valid, (b) distinguish the suitability of imputation for addressing traditional sexism from oppositional sexism, and (c) offer recommendations to minimize harm, maximize benefit, and guide practitioners and researchers on navigating this bind across different contexts.

2 Background

2.1 Demographic Data Imputation

Demographic data is often unavailable, for several reasons, including legal barriers to collection. For example, in the United States, the Equal Credit Opportunity Act originally restricted lenders' ability to collect demographic data as a way to *prevent* intentional discrimination by malicious decision-makers. Similarly, France, along with many other European countries, does not collect racial statistics or codify racial categories as part of a policy of race-blindness (Simon 2015; Escafré-Dublet and Simon 2011). In other cases where data is collected, sharing data between different government agencies may be limited for privacy reasons (King et al. 2023). Privacy may also be particularly relevant to more sensitive attributes; detecting discrimination on the basis of sexual orientation requires collecting information that some may be (reasonably) hesitant to disclose, so optional data collection suffers from selection bias (Tomasev et al. 2021). As a result, in many practical settings, disparity measurement is hampered by a lack of demographic data.

A substantial literature thus examines the issue of algorithmic fairness or disparity estimation without fully known labels. However, without any assumptions or demographic information at all, problems of algorithmic fairness are largely intractable. Instead, researchers assume noisy, imperfect, or partially available estimates of demographic data. In practice, these estimates often come from predictive models. These estimates have been studied both for their statistical properties and downstream fairness implications (Chen et al. 2019; Dong et al. 2025) as well as used for the explicit purposes of assessing disparities (Chen et al. 2019; McCartan et al. 2023; Kwegyir-Aggrey et al. 2024) and training debiased machine learning models (Wang et al. 2020; McCartan et al. 2023; Elzayn et al. 2024). Notably, Bayesian Improved Surname Geocoding (BISG) (Elliott et al. 2009), a naive Bayes model of race using Census name and location data, has been used to motivate real, anti-discrimination regulation: in 2013, the United States Consumer Financial Protection Bureau (CFPB) used imputation to detect racial disparities in auto lending (Consumer Financial Protection Bureau 2014). Other scholars have used parental country of origin in the French context as a proxy for race and ethnicity (Govind, Santini, and Derenoncourt 2025).

2.2 Distinguishing Imputation and Prediction

We have not yet differentiated gender prediction and gender imputation. Here, we present our definitions, where gender imputation is a subset of gender prediction.

We define *gender prediction* as inferring gender from input(s). In machine learning terms, where a predictive model, f , is characterized by $f(X) = Y$, prediction treats gender as Y , assuming that Y is not known a priori. We define *gender imputation* as a subset of prediction, distinguished by its *application* and *interpretation*. Prediction falls under the category of imputation if its application is a specific, limited end, namely to audit or describe existing structures or processes, and its interpretation is used to draw descriptive conclusions in the aggregate, or motivate action at the structural level. Our definitions draw from literatures that use imputation to refer to disparity estimation techniques which use predicted demographic values, such as BISG (Imai, Olivella, and Rosenman 2022; Conderino et al. 2025; Adjaye-Gbewonyo et al. 2014; Greengard and Gelman 2024; Haas et al. 2019; Grundmeier et al. 2015; Derose et al. 2013); to our knowledge, no other body of literature consistently uses another term to describe this practice.³

We now provide some clarifications and observations along with these definitions. First, our definition of gender prediction encompasses a broad range of use cases, tasks, and contexts. It covers predicting gender for synthetic or hypothetical persons, such as AI-generated images and text, and gendering non-human objects or entities. It may also include tasks not explicitly framed as a gender prediction but used to make observations about gender. For instance, values such as unsupervised cluster membership, or a predicted proxy quantity Z (e.g., hair length in images), can be treated as a substitute for gender to draw normative conclusions.⁴ This is because, e.g., substituting long hair for femininity implicitly constructs a gender prediction function. Second, our definition of imputation can have fuzzy boundaries. A technique developed for imputation can be deployed on general prediction tasks: for example, BISG (Elliott et al. 2009) was developed to study and mitigate racial disparities in healthcare, but can be used for individualized ad targeting.

Going forward, we refer to the general act of inferring gender as “gender prediction,” and “gender imputation” when specifically focusing on imputation settings, with the understanding that any illegitimacy concerns that affect prediction also affect imputation. However, the subset of prediction cases that count as imputation access a set of possible benefits (e.g., disparity measurement).⁵

³Other literatures (Bertsimas, Pawlowski, and Zhuo 2018; Amblar, Omar, and Royston 2007) more narrowly define imputation as a semi-supervised setting where ground truth gender labels are partially available, but we do not make this specific distinction.

⁴Characterizing when a proxy can be used to draw normative conclusions is a complex question (Hu and Kohler-Hausmann 2020; Hu 2023; Kohler-Hausmann 2018). However most antidiscrimination law is written based on prespecified identity categories (Arneson 2006; Koppelman 2006). As a result, theories of algorithmic fairness and indirect discrimination are often precisely concerned with connecting proxies to identity.

⁵This does not preclude the possibility of other gender predic-

3 Gender Prediction Is Illegitimate

In this section, we give background on gender and legitimacy, then draw from a long literature of critical scholars to explain why gender prediction is illegitimate. This illegitimacy is serious, as it contributes to political and technological systems that enable or directly perpetrate violence, especially against transgender and nonbinary people.

3.1 Definitions of Gender

Gender can be conceptualized in many possible ways, with multiple aspects (e.g., bodily attributes, gender identity, perceived gender, and gender roles) and interactions weaved into intersectional experiences (e.g., differences between White and Black womanhood) (Keyes, May, and Carrell 2021). While broadly understood as a social construct⁶ (Hacking 1999), different scholars characterize gender very differently. Debates in feminist theory about defining womanhood have spanned decades (Hale 1996; Haslanger 2012a; Jenkins 2016), with formulations as a social position (Haslanger 2012a), a performance (Butler 2002), a “routine accomplishment embedded in everyday interaction,” (West and Zimmerman 1987), and a subjective identity (Stryker 2013; Jenkins 2016).

Gender identity (i.e., self-identified gender)⁷ is one particularly important aspect of gender: transgender advocates argue that prioritizing other facets of gender can lead to excluding trans women from counting as women (Jenkins 2016). However, despite both the diversity of formulations and the importance of identity, Scheuerman, Paul, and Brubaker (2019) show that in computer vision, gender prediction and facial analysis systems de facto reduce gender down to gender presentation. Everyday understandings of gender in dominant cultures are likewise flawed, as they generally center on bodily sex and lack awareness of nonbinary genders and transgender people.

Addressing the multiplicity of gender to mitigate the harms of a single, exclusionary definition requires a degree of definitional flexibility. Just as Hanna et al. (2020) argued for race that the use case determines the relevant meaning of race in that context (e.g., perceived race in some forms of discrimination such as racial profiling, but self-identified race in public health statistics and political mobilization), the same is true for gender. To formulate prop-

tion applications outside of imputation having benefits, such as being useful to Blind or low-vision individuals (Bennett et al. 2021), helping transgender people achieve goals around passing (Chong et al. 2021) or affirm them when they do (Scheuerman, Paul, and Brubaker 2019). These benefits also require critical evaluation, but are beyond the scope of our work.

⁶Social constructs are assuredly “real” (Elder-Vass 2012). Rather, social constructs such as gender are shaped by different societal forces; they can be constructed *differently* in different contexts and changed by trends and events, as opposed to being uncontestable. Haslanger (2012b) argues that proposing and adopting definitions of social constructs can and should be driven by normative, ethical, and ultimately political, goals.

⁷There are differing formal accounts of gender identity, but our work will rely on a colloquial definition as “a sense of oneself as a man, woman, or some other gender” (Jenkins 2018).

erly trans-inclusive definitions, trans philosopher Talia Mae Bettcher (Bettcher 2013, 2014) suggests a multiple-meaning view that explicitly distinguishes between dominant (i.e., mainstream, trans-exclusionary) and resistant (i.e., trans-centered) definitions of what it means to, e.g., be a woman. While we do not adopt Bettcher’s definitions exactly, we take this multiple-meaning separation of dominant and resistant definitions as a way to accept the primacy of gender identity while also recognizing the relevance of different aspects or definitions of gender to different settings such as discrimination, cultural norms, or structural inequality. In this paper, when we speak of genders such as “man” or “woman”, we refer to gender identity by default. When using other formulations of gender relevant to particular ends, we refer to these as specific aspects of gender (e.g., “perceived gender”, “gender expression”, or “gender as social position”); we provide a table describing these terms, the contexts to which they may be relevant, and how they are observed in App. C.

3.2 Definitions of Legitimacy

Definitions of legitimacy (Peter 2023; Beetham 2013; Kratochwil 2006) can be descriptive (e.g., examining if an institution, norm, or practice is accepted by a social group (Suchman 1995; Gerth and Mills 1946; Zelditch 2001)) or prescriptive (e.g., normative criteria for whether an entity or practice is acceptable (Stillman 1974; Adams 2022)). For our purposes, we focus on a prescriptive notion of legitimacy, with an emphasis on aligning with trans-inclusive gender justice. This necessitates diverging from majoritarian perspectives on legitimacy; as most of the world is presently cisgender, our framework must be capable of rejecting trans-exclusionary institutions and practices that remain widely accepted. Instead of providing a positive definition for legitimacy, we anchor on a criterion for determining if gender prediction is *not* legitimate: a gender prediction is illegitimate if it restricts individuals’ agency and capacity to self-determine their gender. We emphasize that this notion of self-determination is not merely internal, and includes aligning external expression and social recognition to respect identity. When a prediction is illegitimate, it should not be given governing power or social authority (Peter 2023); it has no standing to assert a person’s gender, override their own account of their gender, or serve as evidence for a contradicting claim. Doing so normalizes systems that harm transgender and nonbinary people who seek to exercise self-determination. Our definition, however, does not mean that an illegitimate prediction is never permissible to use. Rather, such a restriction on gender self-determination must be warranted by the specific purpose it serves, and only inasmuch as that purpose requires.⁸ As the next section demonstrates, every gender prediction is illegitimate to some degree. In this work, we use the language of legitimacy, instead of transphobia or cissexism, to emphasize the role of this au-

thority in gender self-determination.

3.3 Why Gender Prediction Is Illegitimate

While there are many critiques of gender prediction, we distill them into three kinds: against the act of predicting gender, against the reinforcement of gendered norms, and against the classification system of gender labels. While prior works may not originally formulate their critiques in terms of legitimacy, or these kinds, we re-interpret them here to better disentangle their validity implications, which we discuss in the next section. Our discussion centers prototypical cases of machine learning predictions of gender identity of individual humans, which we show extend to other aspects of gender and less conventional applications.

Gender Predictions The process of assigning a gender label is illegitimate, especially for individuals. First, automatically assigning and policing a person’s gender further entrenches administrative harms: Katyal and Jung (2021) link gender, the state, and surveillance capitalism, where predicting gender further empowers government and corporate institutional control (e.g., in suppressing LGBTQ+ content on social media). Costanza-Chock (2020) names the privacy-violating experience of being searched by the TSA as a trans woman, and Keyes (2018) warns of the possibility of gender prediction being used in bathrooms, with potentially violent enforcement. Of course, not all applications of gender prediction actively enable violence, but all applications must assess these kinds of risks.

Second, the actual process of prediction is harmful. Misgendering, or erroneously assigning a gender to a person, directly harms (Kukla and Lance 2023; Howansky et al. 2022; Kapusta 2016; McLemore 2018; Matsuno et al. 2024) individuals, especially (but not limited to) transgender people. As a second-order effect, machine learning exacerbates this problem: the seeming objectivity of technology normalizes gender predictions by humans, and can enable intentional transphobia. Note that misgendering is harmful even when referring to aspects of gender such as gender expression; telling someone that they do not look like their gender identity is generally still a dignitary insult. Less rigid categories or labels may lead to less severe misgendering acts (e.g., describing someone’s appearance as “more masculine” is a less definitive claim), but remain inherently presumptive.

Third, even if a prediction is accurate and does not misgender, it hinges on the assumption that gender is something that can be predicted. This presupposes a definition of gender that decenters gender identity, conflicting with trans-centered principles of gender self-determination. This disrespects the inherent, first-person authority people have over their gender (Bettcher and Shrager 2009). Similarly, Keyes (2018) categorically rejects the idea that gender prediction is compatible with trans identity, writing, “a trans-inclusive system for non-consensually defining someone’s gender is a contradiction in terms.” Even when claims are explicitly scoped but about other aspects of gender (e.g., perceived gender), this framing can still privilege externally assigned perceptions over individuals’ own gender identities.

⁸Consider an analogy to Thomson (1990)’s account of rights: any restriction on rights is an infringement that requires justification, but only an infringement that lacks one is a violation. We borrow this structure without adopting her specific position on rights.

Gender Norms Gender prediction reinforces gendered norms and associations between observable features and a person's gender. Technologies using biometric data, such as facial analysis, are built on especially harmful bioessentialist norms (i.e., they equate gender and sex) (Schumann et al. 2021). While other norms such as gendered names are less egregious, gendered language associations can still cause harm (Gautam et al. 2024). These associations make it difficult for a person to have their gender identity respected, or align others' perceptions of their gender with their identity. Moreover, associations can vary between cultural context and race: Buolamwini and Gebre (2018) found that gender prediction using facial analysis is less accurate for Black women.

Moreover, while humans can also predict gender via harmful gender norms, using technology introduces new harms. Algorithmic gender prediction changes the fundamental dynamics of gender performativity, which makes gendered norms more difficult to contest and harder to change. As Butler (2002) describes, gender is a feedback loop of performance. Every act of gender is performed in response to the expectations that gender itself creates, projecting norms from the past onto the present. However, while human interaction is full of opportunities to shift and contest these gendered expectations, algorithmic systems are fixed by default, and model updates occur at the discretion of the developer. To give an example, Chong et al. (2021) describes cases of transgender people using automated gender prediction to see if they pass as their gender (i.e., appear as their preferred gender presentation). These attempts are tied to a fixed decision boundary between man and woman. Transgender individuals must force themselves to fit an algorithmic standard to be perceived as a particular gender, instead of being empowered to change such standards.

Gender Labels Gender prediction reinforces gender as a restrictive system of classification. Whether in bureaucracy or in databases, many gender schemas are insufficiently expressive, and either erase genders other than "man" and "woman," or homogenize them (i.e., grouping genderqueer, nonbinary, and gender non-conforming together). Those who do not fit into these classifications end up in residual categories that silence them (Star and Bowker 2007; Bowker and Star 2000). Trans men and women, too, face barriers to changing their categorization in this system, even if the categories of "man" and "woman" exist.

Using these classifications in systems of governance harms vulnerable individuals. Critical work by transgender advocate Spade (2015) coins *administrative violence* to demonstrate how the administrative state enables and enacts violence on transgender and nonbinary people. Hoffmann (2021) extends this analysis to technological systems with *data violence*, as data-driven systems reinforce the discursive structures that normalize violence against those on the margins. Note that labels not based on identity, or not applied to real and whole individual people, still reinforce classification systems that limit gender self-determination (e.g., the arbitrariness of attempting to censor "female-presenting nipples" (Pilipets and Paasonen 2022)).

4 Gender Imputation May Yield Valid Disparity Measurements

In the previous section, we provided an overview of reasons previous scholars have argued for the illegitimacy of gender prediction, and thus the harms it causes. We wholly agree with those arguments, but here argue that in the *imputation* case of predictions, the illegitimacy of the imputation does not necessarily indict the validity of measurements based on those imputations. Moreover, we argue that some forms of discrimination are inherently built on different aspects of gender (e.g., those against the perceived gender of individuals), and can by definition be imputed and inferred. Certainly, this does not mean that *all* gender imputation is valid, and many existing applications fall critically short: we provide recommendations in these cases.

4.1 Measurement Validity

Validity allows us to determine what justified claims can be made from a particular measurement (Cronbach and Meehl 1955; Jacobs and Wallach 2021). Adcock and Collier (2001)'s framework distinguishes a *background concept*, encapsulating different notions of disparity (e.g., unequal treatment, unequal opportunity, unequal representation) across different meanings of gender (including those in Sec. 3.1), from a *systematized concept*, the single specific formulation or definition of gender disparity that is the focus of a measurement (e.g., "the wage gap in feminized occupations in the United States"). Wallach et al. (2025) apply this framework to the evaluation of generative AI models and draw on Messick (1993) in suggesting various criteria of validity be applied to examine a measurement. For brevity, we touch on three criteria here that are especially salient. Content validity requires that measurements draw upon the appropriate observed properties and dimensions that make up the concept. Convergent validity examines if a measurement correlates with other established validated measurements. This provides an empirical way of validating disparity measurements that use imputed gender: comparing them to other measures of discrimination, such as in settings where self-reported gender is available. Lastly, consequential validity, which relates to legitimacy, requires that we weigh the consequences of how a measurement shapes the quantities we study, the conclusions we draw, and ultimately, the world.

Of course, making any judgments weighing harms and benefits requires showing that disparity measurements using imputed gender can be valid enough to achieve benefits, which requires careful consideration and design. As Wallach et al. (2025) note, machine learning researchers and practitioners rarely provide sufficiently precise specifications for the systematized concept of interest. Naïve approaches to gender imputation, via conventional prediction models, will almost necessarily be invalid. To understand the validity of gender disparity measurements based on gender prediction, we consider the conceptualization and operationalization steps of the measurement process.

4.2 Conceptualizing Gender for Measuring Discrimination

Conceptualizing (i.e., “systematizing,” in Adcock and Collier (2001); Wallach et al. (2025)’s terms) gender in the fairness context requires intentionally narrowing its scope, as fairness and gender are both essentially contested concepts (Gallie 1955) that require narrowly scoped *pragmatic measurements* (Hand 2004). As we have shown in Sec. 3.1, gender has many different formulations. Similarly, there are many definitions and disagreements about what constitutes discrimination (Hellman 2008; Kohler-Hausmann 2018; Alexander 1992). Just as Hanna et al. (2020) observe with race, we take the position that an effective conceptualization must consider what aspects of fairness are relevant to gender and what aspects of gender are relevant to fairness. Both are dependent on the input data available and understandings of the context.

Conceptualizing Fairness Debates about algorithmic fairness measures arise from differences in substantive value judgments about fairness and justice. No one measurement or concept can capture *all* notions of fairness (Blodgett et al. 2020); Jacobs and Wallach (2021) suggest that we need only be precise about which aspects or purposes motivate us and recognize when they are appropriate. For our purposes, we chiefly focus on disparity as a basic quantitative measure frequently used in the algorithmic fairness and audit literature.

Gender discrimination has multiple dimensions. Trans author Julia Serano distinguishes between two kinds of discrimination central to her experiences as a trans woman: she coins *traditional sexism* to refer to sexism that privileges masculinity over femininity and targets those who are female or feminine, and *oppositional sexism* for sexism that includes transphobia, homophobia, and cissexism and targets gender deviance (Serano 2007). Further, these two kinds of discrimination are not mutually exclusive, and can compound into what she terms *trans-misogyny*; as Serano writes, “I am not dismissed for merely failing to live up to binary gender norms, but for expressing my own femaleness and femininity.” Traditional sexism is a core component of the discrimination affecting some of the most marginalized members of the trans and nonbinary community.

Those concerned with oppositional sexism may be critical, and justifiably so, of those concerned with traditional sexism who employ gender imputation, since that practice will invariably fail to respect transgender and nonbinary identity, likely fail to detect discrimination against nonbinary people, and thus implicitly privilege one form of discrimination over the other. Imputation always has oppositional sexist harms, and we agree that gender imputation, even when valid, should face heavy scrutiny. However, allowing for the *possibility* of imputation yielding valid measurements, without claiming it is legitimate, makes addressing traditional sexism, including instances of trans-misogynistic discrimination, more tractable. Thus, while we primarily discuss imputation’s suitability for measuring traditional sexism, we hope that disentangling traditional from oppositional sexism also highlights how the latter has been obscured by generic accounts of sexism. We proceed with

the understanding that both forms of sexism are important, and that the illegitimacy of imputation is a reason that researchers should develop different measurements and techniques to study oppositional sexism.

Conceptualizing Gender Note that the validity of a disparity measurement using imputed gender differs from the validity of *imputed gender as a measurement itself*, and the validity of one does not inherently determine the other. However, they are in many cases intertwined: for example, measuring discrimination that causes representational harms may straightforwardly involve counting the number of women, or people perceived as women, in a photo.

For this reason, we describe how to choose an appropriate conceptualization of gender relevant to a particular measurement of discrimination. Drawing from Hanna et al. (2020), the conception of gender relevant for discrimination is often different depending on the context. For instance, for hiring discrimination in an interview, visually perceived gender is relevant. For hiring discrimination in resume screening, perceived gender from name is relevant. In both cases, self-identified gender identity is less relevant, though of course it bears on presentation through performance.

By conceptualizing gender under these more measurable formulations (e.g., perceived gender, gendered social position) compared to gender identity, which cannot be predicted, we are better suited to make less illegitimate claims and make more valid measurements. Of course, these more measurable formulations are distinct from and can conflict with gender identity, and such conflict is itself a harm for the reasons listed in Sec. 3.3. Having a feminine perceived gender does not make someone a woman. Distinguishing these formulations, and what an imputed label implies, minimizes the illegitimacy of making a prediction and specifies a more validly measurable concept.

In fact, these other conceptualizations of gender not defined by self-identification can better align with our conceptualization of discrimination and lead to greater validity. For example, misogynistic street harassment generally occurs without regard for a person’s gender identity; a nonbinary person misperceived as a woman might still experience cat-calling. In such a case, being perceived⁹ or socially situated as a woman subjects someone to traditional sexism. Moreover, people perceived as not conforming to gender norms can experience harms that transgender and nonbinary people perceived as conforming do not, regardless of whether they are cisgender (Case 1995), and imputation that “recogniz[es] performative markers of underrepresented genders” could even measure some forms of oppositionally sexist discrimination (Scheuerman, Paul, and Brubaker 2019). Thus, we conclude that imputing conceptualizations of gender other than gender identity can produce disparity measurements that validly systematize traditional sexism, even if imputed

⁹This raises the question: perceived by whom? Often, the perception most relevant for discrimination measures is that of a possibly discriminatory decision-maker. Of course, in most contemporary cases and at scale, discrimination is more complex. Nonetheless, part of what makes structural discrimination what it is is that it is widespread by definition.

values do not reflect individuals' gender in the broad sense.

4.3 Distinguishing Legitimacy from Validity

We clarify here that legitimacy and validity are distinct qualities. Consider a hypothetical algorithm used by an oppressive state that successfully classifies certain individuals into some category in order to treat them poorly: its validity may not be in question, but its authority is illegitimate regardless. We provide a brief two-by-two table distinguishing cases of valid and invalid, and legitimate and illegitimate, gender disparity measurements in App. B.

In examining the validity of disparity measurements that use imputed gender, we make the case that gender imputation is inherently illegitimate, but not in all cases invalid. Of course, gender imputation is *often* invalid for the same reasons that it is illegitimate. For example, classification systems of gender, alongside being harmful and violent, are simply *inaccurate*; gender norms are often far less stable and inherent than model designers assume; and nonbinary people cannot be accurately classified (Scheuerman, Paul, and Brubaker 2019; Keyes 2018).

However, there are cases where a gender imputation may hold some validity. For example, we believe that the experience of feeling like a gender minority in computer science spaces is justified, and grounded in reality, even without knowing others' gender identities. These feelings are grounded in aspects of gender such as perception and social position that are by definition constructed by inferences; rejecting these feelings inherently dismisses these experiences. Of course, such an experience does not justify misgendering others or void harms from doing so. To reconcile these conflicting intuitions, we distinguish illegitimately assuming others' genders from the feeling of "not seeing people who look like me," and distinguish their relevance to the oppositional sexist harm of misgendering from the traditional sexist harm of a misogynistic culture.

Furthermore, taking the position that gender imputation is always invalid may enable even greater harm. Boyd and Sarathy (2022) point to cases where the correctness of statistical procedure becomes a political weapon — used by those unconcerned with the wellbeing of transgender and nonbinary populations. For example, the CFPB's use of imputed race (Consumer Financial Protection Bureau 2014) was accused of being statistically invalid (Baines and Courchane 2014) by corporations and political actors who sought to curtail their regulatory powers and weaken the foundations of anti-discrimination law. Similarly, the harms of undetected discrimination may trade off against the harms of using gender imputation when the former is difficult or impossible to quantify by other means. Such judgments can only be made if we allow for the possibility of assessing the validity and therefore effectiveness of imputation. In the next section, we provide recommendations for assessing and improving this effectiveness, and determining when deploying imputation might be warranted in spite of its inherent illegitimacy.

5 Recommendations

We have argued that gender imputation is always illegitimate, and therefore harmful and unsuitable for measuring

oppositional sexism, but may sometimes be valid for measuring discrimination related to traditional sexism. However, while we strongly argue for acknowledging the validity of gender disparity measurements based on imputation, weighing if or when an application of gender imputation is justified is a separate question. Making this judgment requires exploring whether alternative ways of characterizing gender fairness exist, as well as the possibilities of minimizing illegitimacy while maximizing validity and benefit, where the status quo is currently insufficient in both regards. Here, we provide recommendations toward these ends when constructing measurements that use imputed gender.

5.1 Weighing Tradeoffs between Illegitimacy and Validity

We draw on the principle of beneficence from the 1979 Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1979), which established ethical guidelines for research, to suggest that we should consider how to 1) maximize possible benefits and 2) minimize possible harms.

A justified application of gender imputation should bring otherwise unachievable benefits. Note that the validity of a disparity measurement is instrumental to achieving benefits, but not necessarily a benefit in itself. Even validly measuring disparities and showing them to be normatively wrong may not constitute a substantive benefit if such a diagnosis does not lead to downstream awareness, actions, or impacts. An imputation model that is carefully defined and perfectly accurate remains unacceptable if the audited task is normatively trivial. We suggest weighing harms against the consequences of inaction: leaving gender unexamined in settings without self-reported data or other mechanisms for detecting discrimination may also risk severe harm.

Harm minimization, on the other hand, can help to address some legitimacy issues. As imputation will always be to some degree illegitimate, any usage must minimally restrict gender self-determination. A general-use imputation of gender as a whole cannot be justified over using a more narrowly tailored gender concept, and imputation cannot be justified if there are reasonable alternatives for calculating disparity or acquiring gender data (e.g., joining records with self-identified sources (Bennett et al. 2021)). Minimizing any delegitimizing factors also maximizes consequential validity. For instance, making predictions on vulnerable populations and enabling downstream misuse (e.g., by justifying new surveillance infrastructure) are highly illegitimate and harmful uses, as discussed in Sec. 3. We suggest the following application-level recommendations in the rest of this section to maximize benefit and minimize harm when judging if imputing gender to measure disparity is worthwhile.

5.2 Measurements Should Use a Narrowly Scoped Conceptualization of Gender

Examining a specific task, such as detecting gendered algorithmic bias in resumes, requires operationalizing a contextual and narrow conceptualization of gender. An imputation technique here should not predict a general gendered social

position or perceived gender as a static property of a person, but rather that person's gendered social position or perceived gender in that particular setting and set of norms. For example, Haslanger (2012a) argues that a man of color may not be socially positioned as a man when experiencing emasculating racial violence. Similarly, perceived gender depends on name, appearance, voice, occupation, social role, and varies with which properties are observed; for instance, a cisgender woman named "Alex" might be assumed to be a man from her name, but a woman from her appearance.

Strengthening content validity requires accounting for the complexity of these gendered norms and relationships. Moreover, these gendered positions or perceptions that features imply also vary between time, context, and culture. For example, the early days of computing were led by women (Light 1999; Abbate 2012), and marriage comes with different socially positioned consequences in the modern day than in historical settings. Perceptions can also vary; a dataset of names drawn from Spanish speakers will have different gendered associations with the masculine name "Juan" than an English-language dataset of mainland Chinese names, where a common feminine character is romanized the same way.

Imputation techniques must also actually align with this narrow conceptualization by using appropriately defined and collected data and minimizing the scope of claims made about an individual. For instance, combining two name datasets when one is associated with sex assigned at birth and the other with self-reported gender (Mohammad 2020) illegitimately conflates sex and gender and invalidates any downstream disparity measurement. Instead, imputation models should be trained and validated by collecting perceived gender data from names or resumes directly, such as by surveying recruiters. Properly collected data also allows for a consistent interpretation of what an imputation *is*. If an imputation only predicts "perceived gender based on a resume," for example, this treats gender less as a universal constant property of a person, and only says something about an aspect of their gender in a particular context. While still illegitimate, we consider this less illegitimate than attempting to impute a person's gender as a whole.

5.3 Imputations Should Use Appropriate Inputs

Given a properly scoped conceptualization of gender in a particular setting, any imputed values and downstream disparity measurements should use appropriate and valid features that reinforce less illegitimate norms. For example, in social media data, Scheuerman, Paul, and Brubaker (2019) observe that simple image-based gender classifiers ignore captions or visual artifacts that state or imply someone's gender, which both disrespects users' agency and ignores salient information. In other cases, methods may rely on irrelevant features. Meister et al. (2023) find that spurious visual artifacts, such as particular background colors, prove predictive of gender even when the person is entirely removed from the image. Both using inappropriate features and ignoring appropriate features are concerns in facial analysis imputation methods. Using biometric data is not only illegitimate for conflating sex with gender, but

also invalid, as people often perceive gender through other aspects of appearance. While not without its own issues, incorporating performative aspects of gender presentation that individuals have more agency over, such as clothing, makeup, or hairstyle would improve both validity and legitimacy (Scheuerman, Paul, and Brubaker 2019).

5.4 Gender Imputations Should Be Aggregated

Disparity measurements using imputed gender should be examined in the aggregate to avoid drawing conclusions about individuals and calculated with statistically sound methods.

Using imputed gender to calculate aggregate measurements of disparity (e.g., the difference in hiring conditional on gender) can be less illegitimate than using imputed gender values for individual-level decisions. When showing a gender disparity with an aggregate measurement, we minimize reliance on, and authority given to, individual predictions. Intuitively, a social environment such as a computer science conference can broadly feel dominated by men, or devoid of people with a similar gender presentation. This is a discriminatory phenomenon that exists only in the aggregate. While gender is nonetheless perceived, this discriminatory experience is built on a holistic set of observations without hinging on a strong claim about any individual's gender. Aggregate measures are more easily understood as imperfect estimates and leave room for individual-level uncertainty, contestation, and variation.

Moreover, disparity measurements that properly aggregate imputed values can maximize empirically assessable forms of validity. Performing well on these forms of validity requires using best practices in calculating disparity measurements. Chen et al. (2019) and Dong et al. (2025) observe that using raw, continuous model outputs, instead of discretizing predictions into the distinct categories used in self-reported data, as is commonly done, can mitigate statistical biases in aggregate disparity measures.¹⁰ Empirically, these methods lead to results that more closely match results from self-reported data, showing better convergent validity.

6 Case Studies

Lastly, we examine three gender imputation case studies differing in their target population and feature modality.

6.1 Auditing Stereotypes in Image Generation

Background. Generative text-to-image models are known to output images that are not representative nor diverse (Wan et al. 2024). Most commonly for bias evaluations on these models, researchers examine correlations between gender and occupation: generated images of doctors or managers disproportionately look like men, while images of nurses and secretaries look like women (Luccioni et al. 2023; Naik and Nushi 2023). This brings out our conflicting intuitions:

¹⁰Note that these statistical properties rest on a model calibration assumption. Calibration is often used to interpret model outputs as probabilities, but we do not necessarily endorse this interpretation as either philosophically rigorous (Hu 2025) or trans-inclusive. Of course, any mitigation of estimation bias, like calibration itself, can be empirically tested in a given context by gathering real data.

explicitly recognizing this bias in the image generation setting requires being able to articulate what it means to “look like men” — thus motivating gender imputation.

Legitimacy. Since generated images do not reflect real individuals, not all the harms of making a gender prediction apply. Gender prediction models cannot misgender generated images of people, because there is no real person to experience the dignitary harms of misgendering or have authority over a gender identity. Of course, predictions still reify gender norms, labels, and normalize the act of prediction, so many other concerns in Sec. 3.3 apply. Using methods that less directly perform gender classification can also help avoid normalizing these practices. For example, work that builds a reference set of image clusters for different genders, then assigns generated images to those clusters based on embedding distance, sidesteps some implications of direct classification (Luccioni et al. 2023). Still, completely avoiding interpretations of prediction is likely impossible.

Validity. The lack of real individuals in generated images means that only aspects of gender such as perceived gender exist. This aligns with our earlier conceptual distinctions: gender identity is not validly measurable via imputation, but other aspects of gender may be.

6.2 Auditing Gender Disparities in Film

Background. Another setting examines gender representation in film (Jang, Lee, and Lee 2019; Schmidt et al. 2021), by predicting gender frame by frame to identify disparities such as, e.g., differences in total screen time between men and women, often using computer vision.

Legitimacy. Computer vision prediction of gender has many illegitimate dimensions, with few factors that can mitigate this illegitimacy. The established arguments in Sec. 3.3 apply here, although some factors are worth particular notice. Unlike the case of generated images, concerns about misgendering harms apply in full force. As a particularly notable harm, two works that perform this analysis chiefly use face-based gender prediction models, which rely on particularly illegitimate gender norms (Jang, Lee, and Lee 2019; Schmidt et al. 2021). And, while Schmidt et al. (2021) includes an “androgynous” classification label, this may not substantively address issues with gender labels; their “androgynous” label confusingly refers to “either multiple genders on one screen or uncertainty by the model,” with no substantive discussion of gender nonconformity or the implications of ambiguity.

Validity. The validity of the relevant disparity measures here is closely tied to the validity of the imputation method itself, which may be more limited because of the previously discussed issues with facial prediction. While visually perceived gender is relevant and valuable for this analysis, audiences generally see far more than only a face, and perceive gender from a much more expansive set of cues (e.g., the sound of a voice, the gender of address used in dialogue) than still frames. Arguably, with consideration to the ethical concerns of facial recognition technologies (Stark 2019), directly using facial recognition to identify specific actors and matching to their biographies or character descriptions would likely provide more valid measurements.

6.3 Imputing Gender from Names

Background. Another common gender prediction task is inferring gender from names, as given names in many languages are gendered. For example, Vogel and Jurafsky (2012) and Mohammad (2020) use various sources of name-gender associations (e.g., the US Social Security Administration’s list of baby names, US Census publishing, manual data collection) to examine academic authorship and citation. Recent approaches use machine learning for this prediction task, often using social media data (Santamaría and Mihaljević 2018; Hu et al. 2021; Tang et al. 2011; Sebo 2021; Karimi et al. 2016). We also highlight the resume audit experiment design introduced by Bertrand and Mullainathan (2004), who manipulate names on resumes and submit them to real-world job postings. Drawing conclusions about perceived gender, of course, requires gendering names. Gautam et al. (2024) analyze the practice of name-based sociodemographic prediction in detail, discussing concerns in both ethics and validity.

Legitimacy. Gautam et al. (2024) suggest a gradation of ethical permissibility depending on the application. Notably, they find gender imputation on real people unacceptable. Nothing substantively mitigates the illegitimacy of this practice — name-based predictions are uncontested, normalizing, and misgendering, and often use associations based on sex. In addition, Gautam et al. (2024) address predictions not tied to real people — that is, to draw conclusions about gender bias “based on how NLP models handle synthetic names of people assumed to be exclusively female.” Similar to the generative model case in Sec. 6.1, such predictions are marginally acceptable as they do not risk misgendering, but remain concerning as they “entrench hegemonic folk theories of names and people’s identities.”

Validity. Gautam et al. (2024) identify four barriers to validity: 1) biases arise from mapping names exactly onto gender; 2) researchers conflate different aspects of gender; 3) evaluating gender prediction systems requires self-reported data, with poor generalizability; and 4) echoing consequential validity concerns, the construction of a classification system shapes its conclusions. We consider whether these issues can be minimized. First, as discussed in Sec. 5.4, using aggregated measures with continuous estimators can mitigate discretization-based biases. The second and third issues can both be minimized by aligning imputation practices with a clear conceptualization of gender, as discussed in Sec. 5.2. Indeed, mapping names to sex assigned at birth conflates different gendered concepts; however, Gautam et al. (2024) concede that validity is stronger when examining perceived gender, which we further qualify to *perceived gender based on written name*. Scoping the imputed gender concept to “perceived gender based on name” can also warrant recruiting annotators from a fitting background (e.g., job recruiters in anglophone companies) for more robust error evaluation. Lastly, the consequential concerns with classification warrant the need for clarity in conceptualizing discrimination. However, this issue is inherent to *all* studies of discrimination; working within a flawed system does not inherently invalidate findings of disparity, and must be weighed against not identifying discrimination at all (Sec. 5.1).

7 Discussion

Our work has argued that gender imputation is illegitimate but may be used for valid disparity measurements which pertain to traditional sexism. While considering any particular instance requires care and nuance, we provide ways to weigh and evaluate any benefits imputation may bring against its inherently harmful illegitimacy. Our analysis critiques existing imputation practices and resolves conceptual tensions in the normative obligations around gender prediction.

7.1 Takeaways

Thematically, we emphasize two takeaways. First, being precise about gender imputation highlights the pervasiveness of trans-exclusionary assumptions. Taking a moral prohibition against gender prediction seriously requires addressing cases where well-intentioned or epistemically valid motivations exist. Second, there are forms of discrimination that are based on other aspects of gender. While self-identification and gender identity deserve primacy, they do not capture *all* facets of how gender manifests in the world or in individual experience. Consider that many trans people experience misogyny without being women. Engaging with the nuances allows us to identify discrimination (e.g., on the basis of a feminine appearance) without *ascribing* an identity (e.g., on the basis of being a woman).

More practically, our work has implications both for advocates and for algorithmic fairness practitioners. Although many gender prediction use-cases are poorly motivated, harmful, and irredeemably flawed, clarity about why the practice is illegitimate can help set the bounds of what is permissible in gender bias research (such as the guidelines provided by Scheuerman, Paul, and Brubaker (2019) and Gautam et al. (2024)). Furthermore, without precise and careful engagement with trans philosophy and other socially just ends, concerns about misgendering and oppositionally sexist harms will continue to be sidelined as unreasonable or easily outweighed by other concerns. We observe that this lack of clarity has led to practitioners acknowledging these harms in theory but failing to navigate them in practice. Our contributions here provide tools to navigate this tension, highlighting the need for methods to capture oppositional sexism, beyond simply rejecting imputation.

For researchers or practitioners who may use imputation, our recommendations argue that existing practices should be held to a higher standard. By showing that more valid, and less illegitimate, measurements are possible, our work shows that merely making perfunctory acknowledgments that gender prediction is limited and trans-exclusionary is insufficient (Devinney, Björklund, and Björklund 2022). Many unnecessarily illegitimate practices cannot be justified merely because bias assessment provides some benefit. While our recommendations may require more effort, building fair systems takes both imagination and commitment.

7.2 Limitations

While we provide recommendations for improving validity and reducing illegitimacy, we recognize this list is not exhaustive. There is more work to be done in creating imputation

tools that can validly measure discrimination, and reduce harms stemming from their illegitimacy. Future work may also broaden the scope of our initial argument here: we focus on a relatively narrow kind of discrimination (i.e., disparity), for a few aspects of gender. Critically, we echo prior work by Scheuerman, Paul, and Brubaker (2019) calling for developing techniques that can measure oppositional sexism. In clarifying the bind of gender imputation, we hope it brings to light the neglect of evaluations for oppositional sexism. More thoroughly drawing connections to and conclusions about other demographic attributes is also beyond our work. Engaging with more complex theories about how definitions of discrimination and gender are intertwined or more thoroughly challenging the need for demographic data in algorithmic fairness are other normative assumptions made here that can and should be questioned.

8 Conclusion

In this article, we identify and work through conflicting intuitions around algorithmic gender imputation by decomposing them into questions of legitimacy and validity. This separates a normative prohibition of gender prediction due to its illegitimacy from an epistemic incongruity with a trans-inclusive understanding of gender. Our analysis shows that gender imputation is illegitimate, like all gender prediction, but can be useful in describing real forms of discrimination. To reconcile the normative implications around deploying imputation, we provide the means to analyze and minimize the harms inherent to gender imputation while maximizing the potential benefits from a particular context. We recommend practices to navigate this tension by minimizing the harms of illegitimacy and achieving benefits from maximizing validity. Our work charts a path toward identifying discrimination without reifying the harmful assumptions that underlie it. With this, we can better address the core issues in traditional sexism, while also casting light on the overlooked needs of oppositional sexism.

A Table of LGBTQIA+ Terms and Language

Table 1 collates LGBTQIA+ terms used in our paper, as defined by a crowdsourced wiki run by LGBTQIA+ community members.

B Table of Legitimacy and Validity Examples

Table 2 gives a two-by-two table of examples of gender data and gender imputation used for anti-discrimination purposes, for additional clarity. All four example measurements center around measuring different kinds of gender disparities in the employment setting.

C Table of Gender Terms

Table 3 overviews many of the aspects of gender we describe, as well as other terminology useful for distinguishing gender and sex.

Positionality Statement

Both authors of this work are queer researchers embedded in the algorithmic fairness research community. The

Term	Definition
cissexism	A form of sexism, specifically, a way of thought in which only cisgender people are seen as normal or right. Cissexism is harmful to all kinds of transgender people, including nonbinary people.
misgender	To address someone in a way that contradicts their gender identity. This can be accidental, but if intentional, it can be an example of discrimination against transgender people (cissexism).
non-binary gender, nonbinary gender, or nonbinary	An umbrella term for all who don't identify as just female or male. Though there are many kinds of nonbinary gender identities, some people identify as "nonbinary" only.
trans	Short for transgender or transsexual.
transgender	An umbrella term for those with gender identities that don't match the genders they were assigned at birth.
trans-misogyny	Discrimination and hate crimes against transgender women.

Table 1: All terms are directly sourced from the community-run Nonbinary wiki's English glossary of gender terms.

	Valid	Invalid
Legitimate	using self-reported gender identity to measure wage differences	using self-reported gender identity to measure online recruiter discrimination
Illegitimate	imputing perceived gender from scraped online photos to measure online recruiter discrimination	imputing perceived gender from scraped online photos to measure wage differences

Table 2: Examples of different disparity measurements using gender and imputed gender. Imputed values are always illegitimate, while properly collected self-reported data is one example of legitimate data. The valid examples use aspects of gender aligned with the form of discrimination we intend to measure (gender identity corresponding to the gender wage gap, and perceived gender corresponding to treatment by recruiters perceiving candidates), while invalid ones do not. Note that self-reported gender identity, while legitimate, is not the basis on which online gender discrimination may occur, and may lead to less valid measurements.

first author is a genderfluid nonbinary person affected by trans-misogyny, and their perspective in writing this paper is informed by their personal experiences with having their gender presentation perceived differently in different contexts, and the limitations of articulating that experience in sociotechnical accounts of data collection and algorithmic discrimination.

Ethical Considerations Statement

Our work discusses transphobia and trans-exclusionary ideas on technical and philosophical levels, where uncritically repeating harmful lines of reasoning can normalize transphobia. We considered multiple formulations of our argument and ultimately chose this particular framing as one most likely to achieve our goal (addressing the bind we describe in the paper) while minimizing the chance of justifying harmful practices or ideas, which we describe more in the adverse impacts section.

Adverse Impacts Statement

As discussed in our ethical considerations statement, we were concerned with three main adverse impact risks: 1) our work being twisted to justify harmful uses of imputation, 2) contributing to a discourse that decenters trans-inclusive

definitions and aspects of gender, and 3) enabling harmful research practices even when an application is justified (i.e., insufficiently minimizing harm). To mitigate the first, we formulated the argument to be as narrowly tailored as possible: we avoid contesting the ethical injunction against gender prediction by distinguishing legitimacy from validity, we narrowly scope the definition of imputation, and we formulate our arguments to only be relevant to discrimination to minimize how easily they could validate other applications of imputation, alongside explicitly saying in the paper that most imputation cases will still be unjustified. The second we address by directly stating our normative commitments to trans inclusion in our discussion of gender and shaping our argument around them. For the third, we attempted to balance both provocation and practicality with our recommendations. We judged it valuable to provide a mix of recommendations that may be more immediately implemented and those that might require greater resources, so as to not enable practitioners to stop at "good enough;" for this reason, we also do not treat these recommendations as exhaustive.

Generative AI Usage Statement

We used Claude Sonnet 4.5 and Claude Opus 4.5 in our literature review to broaden our search for relevant works and

Gender Term	Description	Relevant Contexts and Outcomes
Gender Identity	Subjective, self-defined, often anchored to gender categories	Individual experiences, gender labels, and gender as a concept, anti-trans discrimination
Gender Norm	Associations between observable characteristics (e.g., appearance, social position) and gender categories	Structural patterns, social interactions, stereotyping
Femininity / Masculinity	Characteristics associated with the gender categories “woman” / “man” via gender norms	Stereotyping, social expectations, characterizing gender labels
Gender Expression	Observable characteristics person exhibits, driven by gender identity and gender norms	Behavioral patterns
Gender Presentation	Observable characteristics a person exhibits in order to be perceived as a particular gender, such as clothing, cosmetics, and tone of voice	Passing, interpersonal interactions, self-image
Perceived Gender	What others observe of a person’s gender, broadly construed	Interpersonal interaction, discrimination
- Visually Perceived Gender	What others observe of a person’s gender based on appearance	Representational harms in visual media, sexual harassment
- Perceived Gender Label	A gender label associated with a perceived gender	Interpersonal ascriptions, misgendering and social categorization, discrimination
- Perceived Femininity / Masculinity	How observations in perceived gender relate to femininity / masculinity	Stereotyping, social expectations, discrimination
Sex	Bodily characteristics such as genitalia, hormones, and genetics	Healthcare disparities, reproductive justice
Gender as a Social Position (Haslanger 2012a)	“How one is viewed, how one is treated, and how one’s life is structured socially, legally, and economically,” focusing on power and hierarchy	Political economy and political advocacy, labor, state governance, the carceral system

Table 3: We summarize the language we use to describe several aspects of gender and sex, broadly construed, drawing heavily from Hanna et al. (2020)’s table of race. We note that this list is not exhaustive, and other authors may use different terms, descriptions or definitions (e.g., Jenkins (2018) offers a rigorous account of gender identity in much greater detail).

format the table in the appendix. We also used Claude Opus 5 for feedback on grammar, wording, and argument structure.

Acknowledgments

We thank Aniket Kriplani, Aurora Zhang, Gena Kim, Meera Desai, Nikhil Garg, Thalia Zhang, Qiaoying Chen, Rui-Jie Yew, Vyoma Raman, members of the Cornell AI Policy and Practice Initiative, participants of the 2025 AI and Human Values workshop organized by the Central NY Humanities Corridor, and several anonymous reviewers for their feedback in developing this paper. We are particularly grateful to Daniel Susser for his persistent encouragement and support.

ED acknowledges that this material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2139899. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s)

and do not necessarily reflect the views of the National Science Foundation. AW acknowledges support from Mastercard and the Survival and Flourishing Fund.

References

- Abbate, J. 2012. *Recoding gender: Women’s changing participation in computing*. Mit Press.
- Adams, N. 2022. The concept of legitimacy. *Canadian Journal of Philosophy*, 52(4): 381–395.
- Adcock, R.; and Collier, D. 2001. Measurement validity: A shared standard for qualitative and quantitative research. *American political science review*, 95(3): 529–546.
- Adjaye-Gbewonyo, D.; Bednarczyk, R. A.; Davis, R. L.; and Omer, S. B. 2014. Using the Bayesian Improved Surname Geocoding Method (BISG) to Create a Working Classification of Race and Ethnicity in a Diverse Managed Care Popu-

- lation: A Validation Study. *Health Services Research*, 49(1): 268–283.
- Alexander, L. 1992. What makes wrongful discrimination wrong? Biases, preferences, stereotypes, and proxies. *University of Pennsylvania Law Review*, 141(1): 149–219.
- Ambler, G.; Omar, R. Z.; and Royston, P. 2007. A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Statistical methods in medical research*, 16(3): 277–298.
- Andrus, M.; Spitzer, E.; Brown, J.; and Xiang, A. 2021. What We Can’t Measure, We Can’t Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 249–260.
- Arneson, R. J. 2006. What is wrongful discrimination. *San Diego L. Rev.*, 43: 775.
- Baines, A. P.; and Courchane, M. J. 2014. Fair Lending: Implications for the Indirect Auto Finance Market.
- Beetham, D. 2013. *The legitimization of power*. Bloomsbury Publishing.
- Bennett, C. L.; Gleason, C.; Scheuerman, M. K.; Bigham, J. P.; Guo, A.; and To, A. 2021. “It’s complicated”: Negotiating accessibility and (mis) representation in image descriptions of race, gender, and disability. In *Proceedings of the 2021 chi conference on human factors in computing systems*, 1–19.
- Bertrand, M.; and Mullainathan, S. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review*, 94(4): 991–1013.
- Bertsimas, D.; Pawlowski, C.; and Zhuo, Y. D. 2018. From predictive methods to missing data imputation: an optimization approach. *Journal of Machine Learning Research*, 18(196): 1–39.
- Bettcher, T. M. 2013. Trans women and the meaning of ‘woman’. In Power, N.; Halwani, R.; and Soble, A., eds., *Philosophy of Sex: Contemporary Readings*, 233–250. Rowman & Littlefield, 6th edition.
- Bettcher, T. M. 2014. Trapped in the wrong theory: Re-thinking trans oppression and resistance. *Signs: Journal of Women in Culture and Society*, 39(2): 383–406.
- Bettcher, T. M.; and Shrager, L. 2009. Trans identities and first-person authority. *You’ve changed: sex reassignment and personal identity*, 1: 98–120.
- Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of “Bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476.
- Bogen, M.; Rieke, A.; and Ahmed, S. 2020. Awareness in practice: tensions in access to sensitive attribute data for antidiscrimination. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 492–500.
- Bowker, G. C.; and Star, S. L. 2000. *Sorting things out: Classification and its consequences*. MIT press.
- Boyd, D.; and Sarathy, J. 2022. Differential perspectives: Epistemic disconnects surrounding the US Census Bureau’s use of differential privacy. *boyd, danah, & Sarathy, Jayshree.(2022). Differential Perspectives: Epistemic Disconnects Surrounding the US Census Bureau’s Use of Differential Privacy. Harvard Data Science Review,(Special Issue 2). https://doi.org/10.1162/99608f92.66882f0e.*
- Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.
- Butler, J. 2002. *Gender trouble*. routledge.
- Case, M. A. 1995. Disaggregating gender from sex and sexual orientation: The effeminate man in the law and feminist jurisprudence. *Yale Law Journal*, 1–105.
- Chen, J.; Kallus, N.; Mao, X.; Svacha, G.; and Udell, M. 2019. Fairness Under Unawareness: Assessing Disparity when Protected Class is Unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* ’19*, 339–348. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.
- Chong, T.; Maudet, N.; Harima, K.; and Igarashi, T. 2021. Exploring a makeup support system for transgender passing based on automatic gender recognition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13.
- Conderino, S.; Divers, J.; Dodson, J. A.; Thorpe, L. E.; Weiner, M. G.; and Adhikari, S. 2025. Evaluating Methods for Imputing Race and Ethnicity in Electronic Health Record Data. *Health Services Research*, e14649.
- Consumer Financial Protection Bureau. 2014. Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity: A Methodology and Assessment. *Washington, DC: CFPB, Summer*.
- Costanza-Chock, S. 2020. *Design justice: Community-led practices to build the worlds we need*. The MIT Press.
- Cronbach, L. J.; and Meehl, P. E. 1955. Construct validity in psychological tests. *Psychological bulletin*, 52(4): 281.
- Derosé, S. F.; Contreras, R.; Coleman, K. J.; Koebnick, C.; and Jacobsen, S. J. 2013. Race and ethnicity data quality and imputation using US Census data in an integrated health system: the Kaiser Permanente Southern California experience. *Medical Care Research and Review*, 70(3): 330–345.
- Devinney, H.; Björklund, J.; and Björklund, H. 2022. Theories of “gender” in nlp bias research. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 2083–2102.
- Dong, E.; Schein, A.; Wang, Y.; and Garg, N. 2025. Addressing discretization-induced bias in demographic prediction. *PNAS nexus*, 4(2): pgaf027.
- Elder-Vass, D. 2012. Towards a realist social constructionism. *Sociologia, problemas e práticas*, (70): 9–24.
- Elliott, M. N.; Morrison, P. A.; Fremont, A.; McCaffrey, D. F.; Pantoja, P.; and Lurie, N. 2009. Using the Census Bureau’s Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities. *Health Services and Outcomes Research Methodology*, 9: 69–83.

- Elzayn, H.; Black, E.; Vossler, P.; Jo, N.; Goldin, J.; and Ho, D. E. 2024. Estimating and implementing conventional fairness metrics with probabilistic protected features. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 161–193. IEEE.
- Escafré-Dublet, A.; and Simon, P. 2011. Ethnic statistics in Europe: The paradox of colorblindness. *European multiculturalisms: Cultural, religious, and ethnic challenges*, 213–37.
- Fry, R. 2025. Gender pay gap in U.S. has narrowed slightly over 2 decades.
- Gallie, W. B. 1955. Essentially contested concepts. In *Proceedings of the Aristotelian society*, volume 56, 167–198. JSTOR.
- Gautam, V.; Subramonian, A.; Lauscher, A.; and Keyes, O. 2024. Stop! in the name of flaws: Disentangling personal names and sociodemographic attributes in NLP. *arXiv preprint arXiv:2405.17159*.
- Gerth, H. H.; and Mills, C. W. 1946. Politics as a Vocation. *From max weber: Essays in Sociology*, 77–128.
- Govind, Y.; Santini, P.; and Derenoncourt, E. 2025. Racial Inequality in France. Technical report, National Bureau of Economic Research.
- Greengard, P.; and Gelman, A. 2024. An Improved BISG for Inferring Race from Surname and Geolocation. *arXiv:2304.09126*.
- Grundmeier, R. W.; Song, L.; Ramos, M. J.; Fiks, A. G.; Elliott, M. N.; Fremont, A.; Pace, W.; Wasserman, R. C.; and Localio, R. 2015. Imputing missing race/ethnicity in pediatric electronic health records: reducing bias with use of US census location and surname data. *Health services research*, 50(4): 946–960.
- Haas, A.; Elliott, M. N.; Dembosky, J. W.; Adams, J. L.; Wilson-Frederick, S. M.; Mallett, J. S.; Gailliot, S.; Haffer, S. C.; and Haviland, A. M. 2019. Imputation of Race/Ethnicity to Enable Measurement of HEDIS Performance by Race/Ethnicity. *Health Services Research*, 54(1): 13–23.
- Hacking, I. 1999. *The social construction of what?* Harvard university press.
- Hale, J. 1996. Are lesbians women? *Hypatia*, 11(2): 94–121.
- Hand, D. J. 2004. Measurement theory and practice. *London: Arnold*.
- Hanna, A.; Denton, E.; Smart, A.; and Smith-Loud, J. 2020. Towards a Critical Race Methodology in Algorithmic Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 501–512.
- Haslanger, S. 2012a. Gender and Race:(What) Are They?(What) Do We Want Them to Be? 1. In *Arguing about science*, 95–116. Routledge.
- Haslanger, S. 2012b. *Resisting reality: Social construction and social critique*. Oxford University Press.
- Hellman, D. 2008. *When is discrimination wrong?* Harvard University Press.
- Hoffmann, A. L. 2021. Terms of inclusion: Data, discourse, violence. *New Media & Society*, 23(12): 3539–3556.
- Howansky, K.; Wittlin, N.; Bonagura, D.; and Cole, S. 2022. Him, her, them, or none: Misgendering and degendering of transgender individuals. *Psychology & Sexuality*, 13(4): 1026–1040.
- Hu, L. 2023. What is “race” in algorithmic discrimination on the basis of race? *Journal of Moral Philosophy*, 21(1-2): 1–26.
- Hu, L. 2025. Does calibration mean what they say it means; or, the reference class problem rises again. *Philosophical Studies*, 1–27.
- Hu, L.; and Kohler-Hausmann, I. 2020. What’s sex got to do with fair machine learning? *arXiv preprint arXiv:2006.01770*.
- Hu, Y.; Hu, C.; Tran, T.; Kasturi, T.; Joseph, E.; and Gillingham, M. 2021. What’s in a name?—gender classification of names with character based machine learning models. *Data Mining and Knowledge Discovery*, 35(4): 1537–1563.
- Imai, K.; Olivella, S.; and Rosenman, E. T. 2022. Addressing Census Data Problems in Race Imputation via Fully Bayesian Improved Surname Geocoding and Name Supplements. *Science Advances*, 8(49): eadc9824.
- Jacobs, A. Z.; and Wallach, H. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 375–385.
- Jang, J. Y.; Lee, S.; and Lee, B. 2019. Quantification of gender representation bias in commercial films based on image analysis. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–29.
- Jenkins, K. 2016. Amelioration and inclusion: Gender identity and the concept of woman. *Ethics*, 126(2): 394–421.
- Jenkins, K. 2018. Toward an account of gender identity. *Ergo*, 5(27): 713–744.
- Kapusta, S. J. 2016. Misgendering and its moral contestability. *Hypatia*, 31(3): 502–519.
- Karimi, F.; Wagner, C.; Lemmerich, F.; Jadidi, M.; and Strohmaier, M. 2016. Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *Proceedings of the 25th International conference companion on World Wide Web*, 53–54.
- Katyal, S. K.; and Jung, J. Y. 2021. The gender panopticon: AI, gender, and design Justice. *UCLA L. Rev.*, 68: 692.
- Keyes, O. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction*, 2(CSCW): 1–22.
- Keyes, O.; May, C.; and Carrell, A. 2021. You keep using that word: Ways of thinking about gender in computing research. *Proceedings of the ACM on human-computer interaction*, 5(CSCW1): 1–23.
- King, J.; Ho, D.; Gupta, A.; Wu, V.; and Webley-Brown, H. 2023. The privacy-bias tradeoff: Data minimization and racial disparity assessments in US government. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 492–505.

- Kohler-Hausmann, I. 2018. Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw. UL Rev.*, 113: 1163.
- Koppelman, A. 2006. Justice for Large Earlobes-A Comment on Richard Arneson's What Is Wrongful Discrimination. *San Diego L. Rev.*, 43: 809.
- Kratochwil, F. 2006. On legitimacy. *International Relations*, 20(3): 302–308.
- Kukla, Q.; and Lance, M. 2023. Telling gender: The pragmatics and ethics of gender ascriptions. *Ergo an Open Access Journal of Philosophy*, 9.
- Kumar, I. E.; Hines, K. E.; and Dickerson, J. P. 2022. Equalizing credit opportunity in algorithms: Aligning algorithmic fairness research with us fair lending regulation. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 357–368.
- Kwegyir-Aggrey, K.; Durvasula, N.; Wang, J.; and Venkatasubramanian, S. 2024. Observing Context Improves Disparity Estimation when Race is Unobserved. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 796–806.
- Light, J. S. 1999. When computers were women. *Technology and culture*, 40(3): 455–483.
- Luccioni, S.; Akiki, C.; Mitchell, M.; and Jernite, Y. 2023. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36: 56338–56351.
- Matsuno, E.; Bricker, N. L.; Collazo, E. N.; Mohr Jr, R.; and Balsam, K. F. 2024. “The default is just going to be getting misgendered”: Minority stress experiences among non-binary adults. *Psychology of Sexual Orientation and Gender Diversity*, 11(2): 202.
- McCartan, C.; Goldin, J.; Ho, D. E.; and Imai, K. 2023. Estimating Racial Disparities when Race is Not Observed. *arXiv preprint arXiv:2303.02580*.
- McLemore, K. A. 2018. A minority stress perspective on transgender individuals' experiences with misgendering. *Stigma and Health*, 3(1): 53.
- Meister, N.; Zhao, D.; Wang, A.; Ramaswamy, V. V.; Fong, R.; and Russakovsky, O. 2023. Gender artifacts in visual datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4837–4848.
- Messick, S. 1993. Foundations of validity: Meaning and consequences in psychological assessment. *Ets research report series*, 1993(2): i–18.
- Mohammad, S. 2020. Gender gap in natural language processing research: Disparities in authorship and citations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7860–7870.
- Naik, R.; and Nushi, B. 2023. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 786–808.
- National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. U.S. Department of Health and Human Services.
- Ovalle, A.; Goyal, P.; Dhamala, J.; Jagers, Z.; Chang, K.-W.; Galstyan, A.; Zemel, R.; and Gupta, R. 2023. “I’m fully who I am”: Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1246–1266.
- Peter, F. 2023. Political Legitimacy.
- Pilipets, E.; and Paasonen, S. 2022. Nipples, memes, and algorithmic failure: NSFW critique of Tumblr censorship. *New media & society*, 24(6): 1459–1480.
- Reid, P. T. 1988. Racism and sexism: Comparisons and conflicts. In *Eliminating racism: Profiles in controversy*, 203–221. Springer.
- Santamaría, L.; and Mihaljević, H. 2018. Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, 4: e156.
- Scheuerman, M. K.; Paul, J. M.; and Brubaker, J. R. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–33.
- Schmidt, T.; El-Keilany, A.; Eger, J.; and Kurek, S. 2021. Exploring computer vision for film analysis: A case study for five canonical movies. In *2nd international conference of the European association for digital humanities (EADH 2021)*.
- Schumann, C.; Ricco, S.; Prabhu, U.; Ferrari, V.; and Pantofaru, C. 2021. A step toward more inclusive people annotations for fairness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 916–925.
- Sebo, P. 2021. Performance of gender detection tools: a comparative study of name-to-gender inference services. *Journal of the Medical Library Association: JMLA*, 109(3): 414.
- Serano, J. 2007. *Whipping girl: A transsexual woman on sexism and the scapegoating of femininity*. Seal Press.
- Simon, P. 2015. The choice of ignorance: The debate on ethnic and racial statistics in France. In *Social Statistics and Ethnic Diversity: Cross-National Perspectives in Classifications and Identity Politics*, 65–87. Springer International Publishing Cham.
- Smith, A.; and Stewart, A. J. 1983. Approaches to studying racism and sexism in Black women's lives. *Journal of Social issues*, 39(3).
- Spade, D. 2015. *Normal life: Administrative violence, critical trans politics, and the limits of law*. Duke University Press.
- Star, S. L.; and Bowker, G. C. 2007. Enacting silence: Residual categories as a challenge for ethics, information systems, and communication. *Ethics and Information Technology*, 9(4): 273–280.
- Stark, L. 2019. Facial recognition is the plutonium of AI. *XRDS: Crossroads, The ACM Magazine for Students*, 25(3): 50–55.

- Stillman, P. G. 1974. The concept of legitimacy. *Polity*, 7(1): 32–56.
- Stryker, S. 2013. (De) subjugated knowledges: An introduction to transgender studies. In *The transgender studies reader*, 1–17. Routledge.
- Suchman, M. C. 1995. Managing legitimacy: Strategic and institutional approaches. *Academy of management review*, 20(3): 571–610.
- Tang, C.; Ross, K.; Saxena, N.; and Chen, R. 2011. What’s in a name: A study of names, gender inference, and gender behavior in facebook. In *International Conference on Database Systems for Advanced Applications*, 344–356. Springer.
- Thomson, J. J. 1990. *The realm of rights*. Harvard University Press.
- Tomasev, N.; McKee, K. R.; Kay, J.; and Mohamed, S. 2021. Fairness for unobserved characteristics: Insights from technological impacts on queer communities. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 254–265.
- Vogel, A.; and Jurafsky, D. 2012. He said, she said: Gender in the ACL anthology. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, 33–41.
- Wallach, H.; Desai, M.; Cooper, A. F.; Wang, A.; Atalla, C.; Barocas, S.; Blodgett, S. L.; Chouldechova, A.; Corvi, E.; Dow, P. A.; et al. 2025. Position: Evaluating generative ai systems is a social science measurement challenge. *arXiv preprint arXiv:2502.00561*.
- Wan, Y.; Subramonian, A.; Ovalle, A.; Lin, Z.; Suvana, A.; Chance, C.; Bansal, H.; Pattichis, R.; and Chang, K.-W. 2024. Survey of bias in text-to-image generation: Definition, evaluation, and mitigation. *arXiv preprint arXiv:2404.01030*.
- Wang, A. 2025. Identities are not Interchangeable: The Problem of Overgeneralization in Fair Machine Learning. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 485–497.
- Wang, S.; Guo, W.; Narasimhan, H.; Cotter, A.; Gupta, M.; and Jordan, M. 2020. Robust optimization for fairness with noisy protected groups. *Advances in neural information processing systems*, 33: 5190–5203.
- West, C.; and Zimmerman, D. H. 1987. Doing gender. *Gender & society*, 1(2): 125–151.
- Zelditch, M. 2001. Theories of legitimacy. *The psychology of legitimacy: Emerging perspectives on ideology, justice, and intergroup relations*, 33.